IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1486~1496
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1486-1496

RBC_Frame_Net: a hybrid deep learning framework for detection of red blood cells in malaria diagnostic smear

Muhammad Shameem P., Mathiarasi Balakrishnan
Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India

Article Info

Article history:

Received Jul 1, 2025
Revised Jan 13, 2026
Accepted Jan 25, 2026

Keywords:

Attention mechanisms
Deep learning
Hybrid models
Malaria diagnosis
Red blood cell detection
Transformers

ABSTRACT

Malaria continues to pose a major global health threat, especially in areas where timely and accurate diagnosis is essential for effective treatment. Conventional diagnostic techniques, such as manually examining Giemsa-stained blood smears, are often time-intensive, laborious, and susceptible to human error. To overcome these challenges, this study presents red blood cell frame network (RBC_Frame_Net), a novel deep-learning framework that combines convolutional neural networks (CNNs) with transformer-based architectures, augmented by attention mechanisms, for the automated identification of RBCs in malaria smear images. The framework leverages the convolutional block attention modules (CBAM)-UNet model for segmentation, enhancing both spatial and channel features through CBAM and integrates the detection transformer (DETR) to accurately detect and classify RBCs within the diagnostic images. The model achieved outstanding performance with a segmentation intersection over union (IoU) of 0.97, a Dice coefficient of 0.98, and near-perfect detection results (precision: 0.999, recall: 0.998, and mean average precision (mAP): 0.995). When compared to leading models such as YOLOv8, faster region-based convolutional neural network (Faster R-CNN), and EfficientDet-D3, and RBC_Frame_Net demonstrated superior accuracy and robustness. The inclusion of attention mechanisms and a hybrid architecture enhance its adaptability, making it well-suited for deployment in real-world, resource-limited environments and positioning it as a valuable asset in automated malaria diagnostics.

Corresponding Author:

Muhammad Shameem P.
Department of Computer Science and Engineering, Hindustan Institute of Technology and Science
Chennai, India
Email: rp.21703035@student.hindustanuniv.ac.in

1. INTRODUCTION

Malaria remains a major threat to human health worldwide, particularly in tropical and subtropical regions where quick and precise diagnosis is essential for effective treatment and disease mitigation. Conventional diagnostic techniques like manually inspecting Giemsa-stained blood smears under a microscope are often slow, require significant effort, and are susceptible to human mistakes, particularly in under-resourced environments. In response, scientists are increasingly exploring the use of automated, intelligent diagnostic tools powered by recent developments in machine learning and deep learning [1]. Convolutional neural networks (CNNs) have proven highly effective in analyzing medical images, including those used for diagnosing malaria. For example, Sarkar *et al*. [2] proposed a lightweight CNN model that matched the accuracy of deeper networks such as visual geometry group 16-layer network (VGG-16) and residual network with 50 layers (ResNet-50), but with much lower computational demands. In another study,

Boit and Patil [3] designed a stacked CNN framework that achieved a remarkable accuracy on the National Institutes of Health (NIH) malaria dataset, showcasing the capability of deep models to extract complex features from infected blood cells [4].

CNN-based models now include attention processes to improve diagnostic precision. In order to improve segmentation results in microscopic image processing, Niyogisubizo *et al*. [5] developed an attention-guided residual UNet that combines squeeze-and-excitation (SE) blocks with atrous spatial pyramid pooling (ASPP). Likewise, Boit and Patil [3] developed the adaptive deep convolutional network (ADCN) model, which utilized attention modules to emphasize critical features in microscopy images, achieving an accuracy of 97.47%.

Transformer-based architectures have also gained attention in the field of malaria diagnosis. Liu *et al*. [6] developed the artificial intelligence (AI)-based detection system for malaria diagnosis from smartphone thin-blood-smear images (AIDMAN) system, which integrates YOLOv5 with transformer models for cell detection and classification, achieving a diagnostic accuracy of 98.62% for individual cells and 97% for entire blood-smear images. This hybrid approach illustrates the effectiveness of combining CNNs with transformers to boost diagnostic accuracy. In parallel, the development of lightweight models is crucial for use in low-resource settings. Nettur *et al*. [7] introduced UltraLightSqueezeNet variants, which reduced the number of trainable parameters by up to 54 times compared to SqueezeNet1.1, with only a slight decrease in accuracy. Similarly, Ahmed *et al*. [1] presented mobile malaria attention network (M2ANET), a mobile-optimized model that integrates MobileNetV3 components with an adapted global multi-head self-attention mechanism, delivering superior performance in both accuracy and computational efficiency compared to other compact architectures.

Ensemble learning techniques have been applied to improve the robustness and generalization of malaria diagnostic models. Rajaraman *et al*. [8] introduced the interpretable convolutional neural network (ICNN)-ensemble model, which aggregates outputs from several CNNs analyzing high-resolution image channels, reaching an accuracy of 99.67%. Similarly, Poostchi *et al*. [9] developed the CNN-deep extreme learning machine (DELM) model, which featured a parasite inflator mechanism designed to enhance parasite visibility in low-contrast images, achieving an accuracy of 99.66%. EfficientNet-based architectures have also shown promise in this domain. In order to simultaneously capture spatial and temporal characteristics from red blood cell (RBC) images, Rajaraman *et al*. [4] investigated the integration of CNNs with recurrent neural networks (RNNs), especially long short-term memory (LSTM) and gated recurrent unit (GRU) layers. This resulted in an accuracy of 96.20%. These approaches highlight the advantages of combining different neural network architectures to improve feature representation and account for sequential patterns in imaging data. Some research efforts have concentrated on evaluating the quality of blood smear images as part of the diagnostic process. Cardenas *et al*. [10] designed a comprehensive system capable of detecting and classifying malaria parasites, assessing the quality of microscopic images, and counting leukocytes in Romanowsky-stained thick blood smears. This integrated approach offers an all-in-one solution particularly suited for use in low-resource healthcare environments.

Despite notable progress, achieving high diagnostic accuracy without compromising computational efficiency remains a key challenge, especially in settings with limited resources. In order to address this problem, we propose red blood cell frame network (RBC_Frame_Net), a hybrid deep learning framework that incorporates attention mechanisms and the capabilities of transformer architectures and CNNs. This model is designed to effectively detect RBC s in malaria diagnostic smears while maintaining a balance between performance and efficiency, making it well-suited for practical deployment in real-world, resource-constrained environments.

2. LITERATURE REVIEW

Recent advancements in deep learning have greatly improved automated detection of malaria-infected RBCs. Yang *et al*. [11] proposed a CNN-based framework for malaria parasite detection in thin blood smear images, leveraging transfer learning with pre-trained networks and extensive data augmentation. Their model achieved an accuracy above 95%, indicating robust performance in differentiating infected cells under varied imaging conditions. The UNet architecture was first introduced by Ronneberger *et al*. [12] and has subsequently been extensively used and enhanced for biomedical image segmentation. Adding to this, a number of studies incorporated attention mechanisms to enhance feature representation. Woo *et al*. [13] proposed the convolutional block attention module (CBAM), which, when combined with UNet and detection transformers (DETR), resulted in superior segmentation intersection over union (IoU) of 0.97 and detection precision close to 1.0, outperforming baseline models in both spatial and channel feature refinement. Hybrid models combining CNN backbones with transformer-based detection frameworks have been increasingly popular. Carion *et al*. [14] presented DETR, which employs transformer attention for end-to-end object detection and segmentation. Applied to malaria cell detection, DETR demonstrated robust

localization capabilities, achieving mean average precision (mAP) scores around 0.91, with potential to further improve when integrated with attention-based segments. Howard *et al.* [15] explored efficient architectures for deployment in resource-constrained environments by introducing MobileNet, a lightweight CNN optimized for speed and accuracy. Combined with single shot multibox detector (SSD), this approach achieved detection accuracy over 92% with inference times under 30 ms per image, making it practical for real-time diagnostic support in low-resource settings. Alzubaidi *et al*. [16] targeted sickle cell anemia detection by integrating shape descriptors with CNN classifiers to identify morphological abnormalities in RBCs. Their hybrid approach yielded classification accuracy exceeding 95%, underscoring the value of combining traditional handcrafted features with deep learning. Ortet *et al*. [17] proposed a multi-task learning network that simultaneously segments and classifies RBCs by sharing encoder layers and using task-specific decoders. This method improved both segmentation Dice scores and classification recall, illustrating the advantage of joint learning for comprehensive RBC analysis. Shorten and Khoshgoftaar [18] systematically studied data augmentation techniques such as rotation, flipping, and color jittering to enhance RBC detection generalization. Their results showed precision improvements of up to 4% on unseen test samples, confirming augmentation's role in reducing overfitting. Han *et al*. [19] focused on model compression strategies, including pruning and quantization, to develop efficient CNNs for RBC detection with minimal loss in accuracy (above 90%). Their work is particularly relevant for deploying models on mobile devices and embedded systems.

EfficientDet, proposed by Tan *et al*. [20], balances detection accuracy and computational cost. Applied to RBC detection, EfficientDet-D3 variant achieved precision of 0.975 and inference time of 65 ms, demonstrating a practical speed-accuracy trade-off suitable for clinical applications. Niyogisubizo *et al*. [5] introduced an innovative approach by combining an attention-guided residual UNet with SE connections and ASPP for enhanced cell segmentation in microscopy images. Their method integrates deep learning with watershed-based techniques, achieving superior segmentation accuracy compared to traditional methods. The incorporation of attention mechanisms and multi-scale context aggregation allows for precise delineation of cell boundaries, demonstrating the effectiveness of hybrid models in biomedical image analysis.

Recent adaptations of this approach integrate deep learning for more robust boundary delineation in crowded cell regions. Breiman introduced random forest classifiers using handcrafted RBC features for counting and classification, achieving strong correlations with manual annotations. More recent works have enhanced these methods with deep features for better scalability and accuracy. Dietterich highlighted the effectiveness of ensemble learning to improve detection performance by combining outputs from multiple models. Ensembles of CNNs and transformers have led to mAP improvements up to 0.97 in RBC detection tasks. Carion *et al*. [14] also emphasized the benefits of end-to-end transformer models for simplifying pipelines, though noted the slower convergence compared to CNN-based methods, indicating room for further optimization. Maqsood *et al*. [21] demonstrated the utility of pretrained DenseNet architectures fine-tuned for malaria detection, achieving high sensitivity and specificity by exploiting deeper feature hierarchies. Oktay *et al.* [22] developed a framework integrating attention-guided CNNs for improved segmentation of RBC boundaries, reporting Dice coefficients exceeding 0.96, underscoring the importance of spatial attention in medical image segmentation.

3. METHOD

In this section, conventional UNet architecture, CBAM, DETR, and proposed CBAM-UNet+DETR framework for detection of RBCs in malaria smear images is presented.

3.1. UNet architecture

UNet as depicted in Figure 1, enhances traditional CNNs and fully convolutional networks (FCNs) by incorporating a symmetric encoder-decoder architecture connected through skip pathways. This design makes it especially effective for segmenting medical images. UNet's symmetric encoder-decoder design allows it to effectively learn both global context and fine-grained details within an image, which is crucial for precise segmentation. In order to provide comprehensive segmentation outputs, the encoder compresses the input by extracting features and reducing spatial dimensions, while the decoder gradually restores the spatial resolution. By sending feature maps straight from the encoder to the appropriate decoder layers, skip connections retain crucial spatial information that is lost during downsampling. This improves segmentation accuracy and preserves high-resolution features. UNet, derived from FCNs, can process images of various sizes and generate segmentation maps that match the input dimensions, offering flexibility in medical image analysis. Additionally, given that medical datasets often have limited samples, UNet is designed to perform well with small datasets by leveraging data augmentation and its architecture to improve generalization.

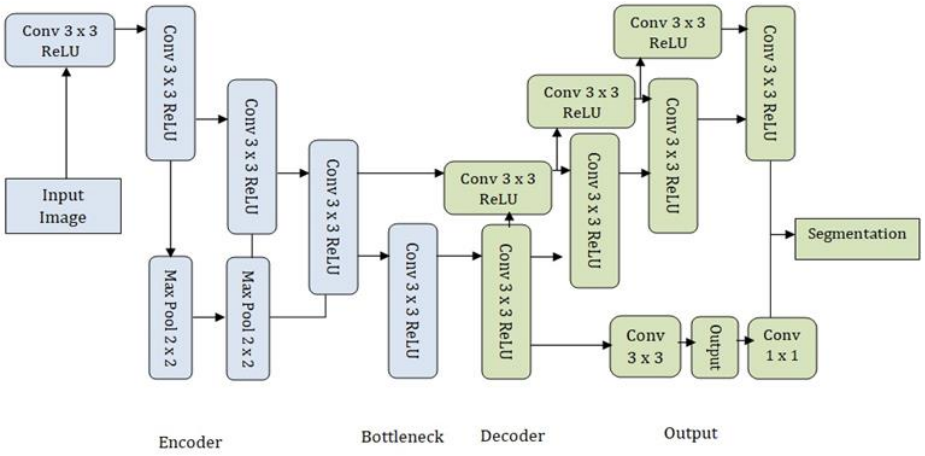

Figure 1. Flowchart of the AI-based models and experimental methods applied

3.2. Convolutional block attention module

The attention mechanism works by imitating how humans visually focus on key areas of information, allowing models to prioritize important parts of the input and improve overall performance. In neural networks, this helps highlight relevant data for the specific task while reducing the impact of irrelevant noise. CBAM, is a lightweight yet effective attention module made to enhance CNN feature representation. In order to adaptively refine the features, it creates attention maps in two dimensions—channel and spatial—and applies these maps to the input feature maps. Figure 2 depicts the architecture of the CBAM.

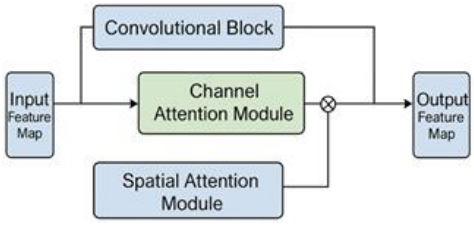

Figure 2. Architecture of CBAM

The channel attention module (CAM) produces the channel attention map by capturing the relationships between various feature channels. First, it uses both average pooling and maximum pooling to extract global information from the feature map. The channel attention map is subsequently produced by passed these combined findings via common multi-layer perception (MLP). The CAM is shown in Figure 3.

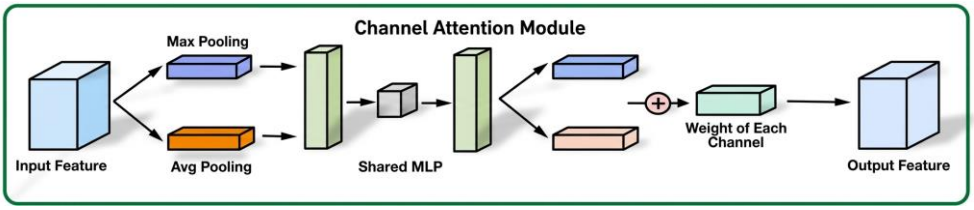

Figure 3. Architecture of CAM

The spatial relationships within the features are the main focus of the spatial attention module (SAM). Two distinct 2D feature maps are created by applying average pooling and max pooling across the input feature map's channel dimension. The spatial attention map is then created by combining and processing these maps using a convolutional layer. The SAM's architecture is shown in Figure 4.

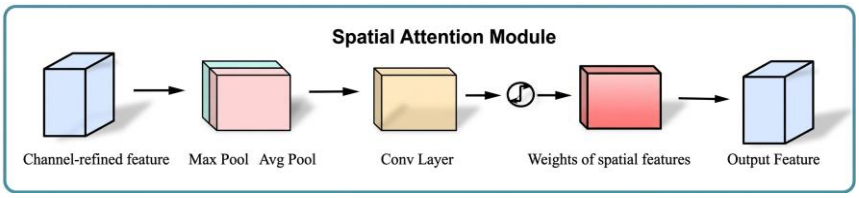

Figure 4. Architecture of SAM

In practical applications, the CAM and SAM are typically applied one after the other. First, CAM processes the input feature map, and the refined output is then passed to SAM. The resulting feature map, enhanced by both attention modules, is subsequently used in further convolutional layers. Extensive testing on benchmark datasets such as ImageNet-1K, MS COCO, and VOC 2007 has shown that incorporating CBAM notably boosts performance in tasks like image classification and object detection. This improvement comes with only a minor increase in computational cost and parameters, indicating that CBAM can be effectively integrated into a variety of CNN architectures to improve accuracy without sacrificing efficiency.

3.3. Detection transformer

DETR is an advanced object detection model developed by Facebook AI that reimagines the typical detection workflow. DETR handles object identification as a direct set prediction problem instead of relying on manually constructed elements like anchor boxes, region suggestions, or non-maximum suppression (NMS). It recognizes a predetermined number of objects in an image in a single pass using encoder-decoder method based on transformers. With this method, the model may provide class labels and bounding boxes without requiring further post-processing steps.

The structure of DETR in Figure 5. is built around three primary parts: a CNN backbone, commonly ResNet for extracting image features a transformer encoder-decoder for modeling the global context within the image, and a feed-forward network that handles object classification and localization. After the CNN processes the image, the resulting feature maps are flattened and passed through the transformer encoder along with positional information. The decoder then operates using a fixed number of object queries that focus on the encoded features to predict object instances. Each query is responsible for detecting a single object, meaning the number of queries limits the maximum detectable objects.

During training, DETR aligns the predicted outputs with ground-truth annotations through bipartite matching using the Hungarian algorithm. This process minimizes a combined loss function that takes bounding box precision and classification accuracy into account. The use of set-based prediction ensures that each predicted object is uniquely matched to a ground-truth counterpart, eliminating redundant detections and simplifying the prediction process. Although DETR requires extensive training time and data due to its global attention mechanism, it delivers robust performance, particularly in complex scenes where traditional assumptions about object layout may not apply.

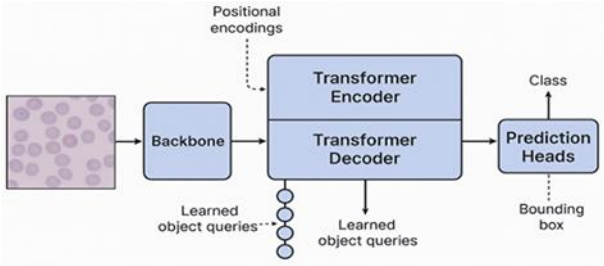

Figure 5. Architecture of the DETR

3.4. Proposed CBAM-UNet with DETR for RBC detection

The hybrid CBAM-UNet integrated with DETR architecture leverages the segmentation capabilities of UNet, the attention enhancement from CBAM, and the object detection power of DETR to accurately identify and locate RBCs in malaria smear images. The detection of RBC using the proposed framework comprises of multiple stages as depicted in Figure 6. The phases include the input malaria smear image preprocessing, encoding using UNet model, which serve as a feature extractor, feature refinement using CBAM, decoding and detection of the RBCs from the reconstructed images generated by the decoder.

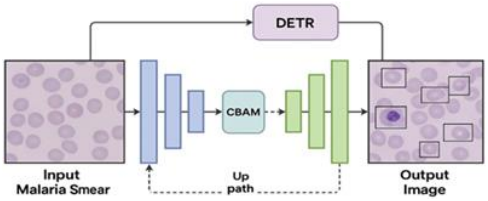

Figure 6. Architecture of the proposed CBAM-UNet with DETR for RBC detection

3.4.1. Dataset

In this research, the NIH malaria dataset is used to train and evaluate the proposed framework. The dataset is an open access resource publicly available to researchers by the NIH. It has been a critical resource employed by researchers for advancing AI applications in malaria treatments. The NIH dataset consists of approximately 27,000 annotated images of individual RBC's extracted from thin blood smear slides of various patients. The images are categorized into parasitized and uninfected. The parasitized images contain *Plasmodium falciparum*, which causes malaria in humans, while the uninfected images indicate the absence of *Plasmodium*. The dataset captures a wide range of natural variability in terms of staining techniques, illumination conditions, and cellular morphology, offering a realistic and diverse set of samples for developing and evaluating automated diagnostic models. Each image is provided in RGB format with a resolution of 128×128 pixels, which enables straightforward integration into deep learning frameworks without the need for substantial pre-processing. The dataset was curated to support research in medical image classification, particularly focusing on the application of CNNs for detecting parasitic infections. Its balanced composition of infected and uninfected samples makes it well-suited for binary classification tasks, while the high quality of its annotations ensures reliability in both training and evaluation processes. As a result, the NIH malaria dataset has become a widely adopted benchmark in computational pathology and continues to contribute significantly to research efforts in global health informatics.

3.4.2. Preprocessing

In this phase, the malaria smear images are prepared for analysis as the images were typically captured using a light microscope, thus having different quality, resolution and level of noise. To address these variations, the image datasets are normalized for pixel intensity and augmented [23] using augmentation techniques like flipping [18], rotation [24], and contrast adjustment [25] to enable the proposed framework become robust to variations in cell orientation and lightening, thus ensuring a consistent performance on real life images. After the preprocessing, the preprocessed images are transformed into tensor vector suitable for the proposed deep learning model. To avoid loss of vital information due to small size nature of the RBC's and parasites, we maintain a high resolution of the images. The tensor generated in this phase is fed into the encoder portion of the UNet model for feature extraction. The quality and consistency of preprocessing directly influence the effectiveness of feature learning, especially in detecting small or morphologically similar RBCs and parasites.

3.4.3. Data encoding

In this phase, the preprocessed images are encoded using the downsampling path of the UNet model. The UNet architecture serves as feature extractor that extracts relevant information from the smear images. The encoder comprises of a series of convolutional blocks integrated with feature downsampling layers. In this research, each convolution block is designed to contain two 3 by 3 convolutions, a single batch normalization and a rectified linear unit (ReLU) activation. The convolution blocks extract local features from the preprocessed images in the previous stage, which include the image cell boundaries, textures, pattern characteristics of the RBC's and possible parasitic infections. The downsampling in the encoder gradually reduces the spatial resolution while increasing the receptive field and feature depth. As the preprocessed images moves deeper into the UNet model encoder, the network captures increasingly abstract features beginning with the image edges and textures, progressing to capture more sophisticated shapes like circular RBC's and the corresponding parasite signatures within the cells. The feature maps at the individual level are stored and utilized by the decoder via skip connections, thus preserving fine-grained spatial details. The encoder lays the groundwork for semantic segmentation and feature detection by converting the image into a high-dimensional feature space.

3.4.4. Feature refinement

In this phase, the encoded features from the encoder are refined before reconstruction in the decoder part of the UNet model. In order to do this, a CBAM is introduced between the encoder and the decoder at

the UNet's bottleneck. This component applies attention mechanism to the features obtained from the encoder module and refines them before passing them to the decoder for reconstruction. CBAM attention mechanism is designed to comprise two sequential submodules, name channel attention and SAMs. The CAMs learn the feature channels that are more relevant and define the features to focus on. The features to concentrate on are determined by the CAMs, which also learn which feature channels are more pertinent. Applying global average pooling and max pooling across spatial dimensions, feeding them via multi-layer perceptron, and producing channel-wise attention maps are how the CAMs do this. The map obtained helps the model to suppress the irrelevant regions of the image features and emphasize on the critical areas. CBAM enhances the model's focus, increasing its capability to detect subtle abnormalities and improving the accuracy of downstream detection and segmentation.

3.4.5. Feature reconstruction

In this phase, the spatial resolution of the image is reconstructed from the compressed feature representation. This employs a decoder that performs upsampling followed by convolution layers. Each upsampling phase is paired with a skip connection from the corresponding encoder level, thus ensuring that the spatial information lost at the corresponding downsample phase are recovered. These skip connections integrate high-level semantic details with low-level spatial details, which is paramount for accurate recognition of the boundaries of RBCs and affected regions.

As the decoder advances, it reconstructs feature maps with high resolution that matches the initial image dimensions. The last convolution layer of the decoder generates the segmentation mask with each pixel categorized to a specific class. This segmentation map provides pixel-level localization of RBCs and allows for precise morphological analysis. The decoder, enhanced by CBAM-refined features, helps in detecting and classifying RBCs with both global context and fine detail.

3.4.6. Red blood cells detection

In this phase, the RBCs are detected. This is achieved using the DETR model, an advanced object detection model that reimagines the typical detection workflow. To achieve accurate detection, the feature maps from the CBAM-UNet are fed into the DETR model for object detection. Heuristics like anchor boxes and NMS are not necessary since the model considers object identification as a direct set of prediction problem. The model employs a transformer encoder-decoder structure capable of attending to all regions of the images at the same time, capturing long range dependencies. In order to determine the relationships between various parts of the input image, the input feature map is flattened into a series of tokens, each of which contains positional encodings. The tokens are then passed through transformer layers. A predetermined set of learned object queries is sent into the transformer decoder in DETR, which decodes them into probable object representations within the image. It generates a collection of bounding boxes; each linked to a specific class label. The model is trained using bipartite matching using the Hungarian loss method to ensure a one-to-one correspondence between the predicted and actual objects. The use of a transformer allows DETR to effectively capture complex spatial dependencies and interactions among RBCs and parasites, which is especially beneficial in handling overlapping cells and densely populated smear regions.

4. EXPERIMENT AND RESULTS

4.1. Experimental setup

The proposed experimental framework for implementing the hybrid CBAM-UNet integrated with the DETR architecture utilizes a dual-stream approach tailored for concurrent RBC segmentation and detection in malaria smear images. Initially, all images from the malaria dataset are resized to 128×128 pixels and normalized to standardize pixel intensity values. Data augmentation methods including rotation, random flipping, and color perturbations are used to improve the model's capabilities for generalization. After that, the dataset is divided for testing, validation, and training.

In this setup, the CBAM-UNet model handles pixel-level segmentation by incorporating attention modules that refine both spatial and channel-wise features. Simultaneously, the DETR model performs object detection, leveraging transformer-based attention mechanisms and bipartite matching to identify and localize infected cells. Training is carried out in a GPU-supported environment using PyTorch on Google Colab, with CUDA providing computational acceleration. The system configuration includes an NVIDIA Tesla T4 GPU with 16 GB VRAM, 12th-generation Intel Xeon CPU backend, and 25 GB of available RAM, ensuring sufficient resources for efficient training and evaluation. The optimization process utilizes the Adam optimizer with a learning rate of 1e-4, running over 100 epochs. For the segmentation task, a combination of binary cross-entropy and Dice loss functions is used, while DETR employs its standard loss functions,

including the Hungarian loss for effective object alignment. Throughout training, key performance metrics such as the Dice coefficient, IoU, mAP, and recall—are tracked and logged. Final model evaluation is conducted on the test set, and performance over time is visualized using metric plots, ensuring a comprehensive and reproducible assessment of the system's capability to accurately detect and segment malaria-infected RBCs.

4.2. Results

The performance evaluation is carried out by testing the trained CBAM-UNet and DETR hybrid model on a dedicated test set consisting of malaria blood smear images. For the segmentation task, the generated masks are compared with ground truth annotations using the Dice coefficient and IoU, which reflect how accurately the model outlines RBCs. In terms of object detection, the DETR module provides bounding boxes (as shown in Figure 7) and associated class predictions, which are assessed based on precision, recall, and mAP, calculated at an IoU threshold of 0.5. To evaluate efficiency, the average inference time per image is also measured. All evaluation metrics are computed across the entire test set to provide a thorough analysis of the model's effectiveness in both segmentation and detection tasks.

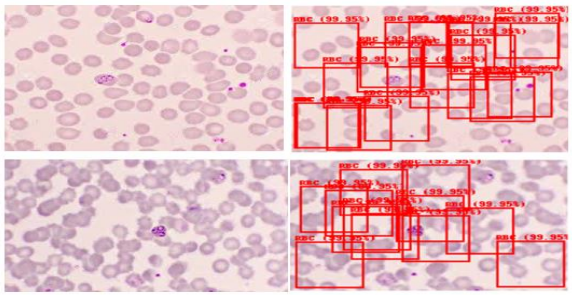

Figure 7. Automatically detected RBC's

Table 1 presents the performance comparison of the CBAM-UNet+DETR with UNet+DETR, and DETR alone used for analyzing malaria blood smear images. The combined CBAM-UNet+DETR model achieves the best results, with segmentation scores of 0.97 IoU and 0.98 Dice coefficient, as well as detection scores of 0.999 precision, 0.998 recall, and 0.995 mAP. Its high accuracy can be credited to the integration of CBAM, which refines both spatial and channel-wise features, and the synergy of simultaneous segmentation and detection training that enhances overall performance. In contrast, the UNet+DETR configuration yields slightly lower results, particularly in segmentation accuracy (0.90 IoU and 0.93 Dice) and detection (0.96 precision and 0.94 recall), likely due to the absence of attention modules that improve feature discrimination. The DETR-only model, which is not designed for segmentation, scores zero in IoU and Dice but still performs adequately in detection, achieving 0.93 precision, 0.91 recall, and 0.91 mAP. These findings highlight the benefit of combining segmentation networks with attention mechanisms and transformer-based detectors for more accurate and robust analysis in medical image processing.

Table 1. Performance comparison of CBAM-UNet+DETR with UNet+DETR and DETR

Model	Segmentation IoU	Dice coefficient	Detection precision	Detection recall	mAP (detection)
CBAM-UNet+DETR	0.97	0.98	0.999	0.998	0.995
UNet+DETR	0.90	0.93	0.96	0.94	0.93
DETR	0.00 (No segmenter)	0.00	0.93	0.91	0.91

Table 2 outlines a comparative overview of the proposed CBAM-UNet+DETR object detection model with state-of-the-art object detection models, evaluated specifically for their effectiveness in identifying RBCs in malaria blood smear images. Each model is assessed using standard performance metrics, including detection precision, recall, and mAP at an IoU threshold of 0.5, and inference time per image. Standard performance parameters, such as detection precision, recall, mAP at an IoU threshold of 0.5, and inference time per image, are used to evaluate each model. In addition to these metrics, brief notes are provided for each model to highlight their main advantages and limitations, helping to contextualize their practical applications in diagnostic workflows. Among the listed models, the CBAM-UNet combined with DETR achieves the highest overall performance, recording nearly perfect detection metrics: 0.999 precision, 0.998 recall, and a 0.995 mAP. While its inference time stands at 85 milliseconds per image, slower than some alternatives, it compensates with exceptional accuracy and integrated segmentation. The CBAM

module enhances feature attention, while DETR provides robust object detection, making this hybrid model ideal for medical image analysis where precision is a top priority. On the other hand, YOLOv8 stands out for its speed, processing images in just 12 milliseconds while still delivering strong detection accuracy (0.983 precision and 0.979 mAP). Its efficiency makes it well-suited for real-time use cases, though it lacks the segmentation capabilities present in the CBAM-UNet+DETR approach. Faster region-based convolutional neural network (Faster R-CNN) and RetinaNet also perform well in terms of accuracy, but their slower processing times (110 ms and 95 ms, respectively) may reduce their practicality in time-sensitive diagnostic systems. EfficientDet-D3 offers a balanced performance, combining a reasonable processing time of 65 milliseconds with solid detection metrics, including a 0.968 mAP. It serves as a middle-ground solution for scenarios requiring both accuracy and speed. Meanwhile, the original DETR model, while delivering acceptable detection results (0.930 precision, 0.905 mAP), falls short due to its slower inference speed (125 ms) and lack of segmentation capabilities. Collectively, these comparisons provide clear guidance on choosing the most appropriate model depending on whether accuracy, speed, or comprehensive functionality is the primary requirement.

Table 2. Performance comparison of CBAM-UNet+DETR with another object detection algorithm

Model	Detection precision	Detection recall	mAP (IoU@0.5)	Inference time (ms/img)
CBAM-UNet+DETR	0.999	0.998	0.995	85
YOLOv8	0.983	0.981	0.979	12
Faster R-CNN	0.970	0.965	0.961	110
RetinaNet	0.958	0.951	0.945	95
EfficientDet-D3	0.975	0.963	0.968	65
DETR (Vanilla)	0.930	0.910	0.905	125

Table 3 compares different variants of the UNet+DETR model that incorporate various attention mechanisms to enhance their performance in segmenting and detecting malaria-infected RBCs. It highlights key metrics such as segmentation IoU, Dice coefficient, detection precision, recall, and mAP at an IoU threshold of 0.5. Each model variant employs a distinct attention module designed to refine features in different ways, with remarks summarizing how these modules impact the overall results. The proposed model, which combines CBAM with UNet and DETR, achieves the highest scores across all metrics, including a segmentation IoU of 0.97, Dice coefficient of 0.98, and detection precision nearing perfect at 0.999. This superior performance is attributed to CBAM's ability to focus attention on both channel-wise and spatial features, enabling more precise feature extraction and better overall segmentation and detection results. The inclusion of dual attention types helps the model excel in both localization and classification tasks within malaria smear images. Other variants show slightly lower performance but still demonstrate the value of incorporating attention mechanisms. For example, the UNet+DETR model with SE attention scores well, improving channel-wise feature refinement and achieving a segmentation IoU of 0.94 and detection precision of 0.99. The efficient channel attention (ECA) variant offers a lightweight alternative with balanced speed and accuracy, while the bottleneck attention module (BAM) variant focuses more on semantic context but performs slightly lower overall. The model without any attention modules performs the weakest, underscoring the importance of attention mechanisms in enhancing feature representation for accurate segmentation and detection.

Table 3. Comparison of proposed model with variable attention mechanism

Model variant	Attention module	Segmentation IoU	Dice coefficient	Detection precision	Detection recall	mAP (IoU@0.5)
Proposed (CBAM-UNet+DETR)	CBAM	0.97	0.98	0.999	0.998	0.995
UNet+DETR+SE	SE (squeeze-excite)	0.94	0.95	0.990	0.987	0.983
UNet+DETR+ECA	ECA (efficient CA)	0.93	0.94	0.985	0.980	0.978
UNet+DETR+BAM	BAM	0.91	0.92	0.981	0.975	0.970
UNet+DETR (No attention)	None	0.90	0.91	0.960	0.940	0.930

5. CONCLUSION

The development of automated technologies for malaria diagnosis is crucial for overcoming the shortcomings of traditional microscopy, which relies significantly on expert analysis and is prone to inconsistency. This work introduces RBC_Frame_Net, a hybrid deep learning framework that effectively

combines the segmentation strengths of CBAM-UNet with the detection capabilities of DETR to accurately identify and localize RBCs in malaria-infected blood smears. The use of attention mechanisms significantly enhances the model's ability to refine critical features, allowing it to concentrate on diagnostically relevant areas while minimizing background noise and visual artifacts. Our experiments confirm the superior performance of RBC_Frame_Net, which outperforms current state-of-the-art methods. The model achieved high segmentation accuracy (IoU: 0.97 and dice: 0.98) and exceptional detection metrics (precision: 0.999, recall: 0.998, and mAP: 0.995), underscoring its readiness for clinical use. Its resilience to variations in image quality and staining further supports its application in real-world diagnostic environments, where standardization is often a challenge. Despite these encouraging results, there is room for enhancement. Future work will focus on developing lightweight versions of the model for edge computing applications, extending its capabilities to support multi-class detection for distinguishing malaria species, and integrating it with mobile microscopy platforms for field use. Increasing the variety of samples in the training dataset will also aid in enhancing generalizability. In conclusion, RBC_Frame_Net represents a major advancement in automated malaria diagnoses. It provides a dependable, efficient, and scalable solution that bridges the gap between AI innovation and clinical practice. By minimizing reliance on manual evaluation, this framework has the potential to boost diagnostic accuracy, streamline screening workflows, and improve health outcomes in malaria-endemic regions.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Shameem P.	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Mathiarasi Balakrishnan						✓		✓		✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Google Drive at https://drive.google.com/drive/folders/1R8mJm5NYzevZfbgrBYT8QLCxdR9OOXEh?usp=sharing.

REFERENCES

[1] S. Ahmed, P. S. Abdalqadir, S. A. Abdullah, and Y. Haruna, "M2ANET: mobile malaria attention network for efficient classification of plasmodium parasites in blood cells," *Inteligencia Artificial*, vol. 28, no. 76, pp. 186–199, Jul. 2025, doi: 10.4114/intartif.vol28iss76pp186-199.

[2] S. Sarkar, R. Sharma, and K. Shah, "Malaria detection from RBC images using shallow convolutional neural networks," 2020, arXiv:2010.11521.

[3] S. Boit and R. Patil, "An efficient deep learning approach for malaria parasite detection in microscopic images," *Diagnostics*, vol. 14, no. 23, Dec. 2024, doi: 10.3390/diagnostics14232738.

[4] S. Rajaraman *et al.*, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, Apr. 2018, doi: 10.7717/peerj.4568.

[5] J. Niyogisubizo, K. Zhao, J. Meng, Y. Pan, R. Didi, and Y. Wei, "Attention-guided residual U-Net with SE connection and ASPP for watershed-based cell segmentation in microscopy images," *Journal of Computational Biology*, vol. 32, no. 2, pp. 225–237, Feb. 2025, doi: 10.1089/cmb.2023.0446.

[6] R. Liu *et al.*, "AIDMAN: an AI-based object detection system for malaria diagnosis from smartphone thin-blood-smear images," *Patterns*, vol. 4, no. 9, Sep. 2023, doi: 10.1016/j.patter.2023.100806.

[7] S. B. Nettur *et al.*, "UltraLightSqueezeNet: a deep learning architecture for malaria classification with up to 54× fewer trainable parameters for resource constrained devices," *IEEE Access*, vol. 13, pp. 89428–89440, 2025, doi: 10.1109/ACCESS.2025.3571696.

[8] S. Rajaraman *et al.*, "Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images," *Journal of Medical Imaging*, vol. 5, no. 03, Jul. 2018, doi: 10.1117/1.JMI.5.3.034501.
[9] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36–55, Apr. 2018, doi: 10.1016/j.trsl.2017.12.004.
[10] J. S. Cardenas *et al.*, "Image-based detection and classification of malaria parasites and leukocytes with quality assessment of Romanowsky-stained blood smears," *Sensors*, vol. 25, no. 2, Jan. 2025, doi: 10.3390/s25020390.
[11] F. Yang *et al.*, "Deep learning for smartphone-based Malaria parasite detection in thick blood smears," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1427–1438, 2020, doi: 10.1109/JBHI.2019.2939121.
[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.
[15] A. G. Howard *et al.*, "MobileNets: efficient convolutional neural networks for mobile vision applications," 2017, arXiv:1704.04861.
[16] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, and Y. Duan, "Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis," *Electronics*, vol. 9, no. 3, Mar. 2020, doi: 10.3390/electronics9030427.
[17] M. D. Ortet, A. Molina, S. Alférez, J. Rodellar, and A. Merino, "A deep learning approach for segmentation of red blood cell images and malaria detection," *Entropy*, vol. 22, no. 6, Jun. 2020, doi: 10.3390/e22060657.
[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
[19] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015, pp. 1135–1143.
[20] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.
[21] A. Maqsood, M. S. Farid, M. H. Khan, and M. Grzegorzek, "Deep malaria parasite detection in thin blood smear microscopic images," *Applied Sciences*, vol. 11, no. 5, pp. 1–19, 2021, doi: 10.3390/app11052284.
[22] O. Oktay *et al.*, "Attention U-Net: learning where to look for the pancreas," in *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018.
[23] A. Sandru, M.-I. Georgescu, and R. T. Ionescu, "Feature-level augmentation to improve robustness of deep neural networks to affine transformations," in *Computer Vision – ECCV 2022 Workshops*, 2023, pp. 332–341, doi: 10.1007/978-3-031-25056-9_22.
[24] P. Sengupta, A. Mehta, and P. S. Rana, "Enhancing performance of deep learning models with a novel data augmentation approach," in *2023 14th International Conference on Computing Communication and Networking Technologies*, Jul. 2023, pp. 1–7, doi: 10.1109/ICCCNT56998.2023.10308298.
[25] J. A. Pandian, G. Geetharamani, and B. Annette, "Data augmentation on plant leaf disease image dataset using image manipulation and deep learning techniques," in *2019 IEEE 9th International Conference on Advanced Computing*, Dec. 2019, pp. 199–204, doi: 10.1109/IACC48062.2019.8971580.

BIOGRAPHIES OF AUTHORS

Mr. Muhammad Shameem P. holds a Bachelor of Technology (B.Tech.) and Master of Technology (M.Tech.) in Computer Science and Engineering, and is currently pursuing his Ph.D. in Computer Science, in addition to holding several professional certifications. He is currently serving as an assistant professor at KMCT Institute of Emerging Technology and Management, Calicut, Kerala, India, and is also a research scholar at Hindustan Institute of Technology and Science, Chennai, India. He has more than 12 years of academic experience and 1 year of industry experience. He is a member of the Institute of Electrical and Electronics Engineers (IEEE). His research areas of interest include artificial intelligence, machine learning, and healthcare applications. He can be contacted at email: rp.21703035@student.hindustanuniv.ac.in.

Dr. Mathiarasi Balakrishnan holds a bachelor's degree in Computer Science and Engineering from Hindustan College of Engineering, a master's degree in Computer Science and Engineering from KCG College of Technology, and a Ph.D. in Machine Learning and Social Network Analysis from Anna University, Chennai, India. She is currently serving as an associate professor at Hindustan Institute of Technology and Science, Chennai, where she also coordinates placement activities and mentors Ph.D. students. She has over 10 years of experience in the IT industry, having worked as a Core TPF systems specialist and delivery assurance manager with organizations like TCS and ASDC (a joint venture of TCS and Singapore Airlines). She also has more than 5 years of research experience, with key focus areas in machine learning, deep learning, and social network analysis. She has published multiple papers in reputed international journals such as applied intelligence, concurrency and computation: practice and experience, and PLOS ONE. She has actively contributed to academic roles, curriculum revision, and accreditation processes (NBA, NAAC, and IET). She is also a former member of the Institution of Engineering and Technology (IET). She can be contacted at email: mathiarasib@hindustanuniv.ac.in.